DESARROLLO Y APLICACIÓN DE LA METODOLOGÍA BAGGING Y ADABOOST PARA LA DETECCIÓN DE PÉRDIDAS NO TÉCNICAS EN EL SISTEMA DE DISTRIBUCIÓN DE LA EMPRESA DE ENERGÍA DE PEREIRA S.A. ESP

ANDRÉS FELIPE GIRALDO DE LOS RÍOS

MAESTRÍA EN INGENIERÍA ELÉCTRICA
FACULTAD DE INGENIERÍAS
UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PEREIRA, ENERO DE 2018

DESARROLLO Y APLICACIÓN DE LA METODOLOGÍA BAGGING Y ADABOOST PARA LA DETECCIÓN DE PÉRDIDAS NO TÉCNICAS EN EL SISTEMA DE DISTRIBUCIÓN DE LA EMPRESA DE ENERGÍA DE PEREIRA S.A. ESP

ANDRÉS FELIPE GIRALDO DE LOS RÍOS

PROYECTO DE GRADO
PARA OPTAR AL TÍTULO DE MAGÍSTER EN INGENIERÍA ELÉCTRICA
LÍNEA DE AUTOMÁTICA Y ELECTRÓNICA

DIRECTOR:

Ph, D. ANDRÉS ESCOBAR MEJÍA

MAESTRÍA EN INGENIERÍA ELÉCTRICA
FACULTAD DE INGENIERÍAS
UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PEREIRA, ENERO DE 2018

Agradecimientos

A mí esposa Juliana y a mí hijo Daniel
quienes fueron mi motor y motivación constante
para culminar este gran reto ...
A mis padres por su apoyo incondicional…
A la Empresa de Energía de Pereira
por su apoyo y facilidades para cumplir este logro…
Al Ing. Andrés Escobar Mejía por su ayuda desde el inicio,
consejos y observaciones…

Contenido

Capítulo 1

Introducción

El suministro del servicio de energía eléctrica a los usuarios residenciales, comerciales, industriales y en general, es producto del proceso de generación, transporte y comercialización del cual hacen parte varios actores. La producción de la energía es tarea de las empresas de generación, su función es emplear recursos como: agua, carbón, energía solar, etc. para convertirlos en energía eléctrica. Desde donde se produce la energía y hasta los diferentes puntos de consumo, se utilizan largas y grandes autopistas que conforman la red de transmisión; de allí se derivan ramales más pequeños o calles que componen la red de distribución y se transforma la energía en niveles adecuados para su comercialización o venta. Esto con el fin de usar electrodomésticos, equipos de oficina, máquinas de producción a nivel industrial o iluminación.

En el proceso de generación, transformación, distribución, se pueden presentar pérdidas desde el punto de generación hasta la entrega final a los usuarios. Dichas pérdidas de energía generan altos costos para los agentes involucrados en la prestación del servicio y los consumidores, impactando a cada uno de estos de la siguiente manera [1]:

1. El comercializador debe pagar al generador y al transmisor el total de la energía que ingresa a su sistema, aunque esta no sea facturada a los usuarios.
2. El distribuidor no recibe el pago por el uso de la infraestructura, asociado con el transporte de la energía que no es facturada.
3. El precio de la energía en Colombia se obtiene mediante un esquema de precio marginal; la generación adicional requerida por la existencia de pérdidas impone un precio marginal mayor el cual es trasladado directamente a los usuarios.
4. Los usuarios pagan un valor adicional al asociado a su consumo mensual, ya que en la tarifa se incluye el costo de pérdidas reconocidas en generación, transmisión y distribución.

La Empresa de Energía de Pereira S.A. E.S.P (EEP S.A. ESP) realiza esfuerzos importantes con el fin de garantizar la reducción en el indicador de pérdidas técnicas y no técnicas en su sistema de distribución. Las pérdidas técnicas son aquellas que son ocasionadas a la operación normal del sistema y consisten en la disipación de potencia en los componentes del sistema eléctrico, tales como líneas de distribución y transformadores. Por su parte, las pérdidas no técnicas son causadas principalmente por la alteración de los equipos de medida (hurto de energía) por parte de los usuarios.

El problema de las pérdidas no técnicas de energía es enfrentado a nivel mundial, con una mayor incidencia en los países en vía de desarrollo, dada la ilegalidad y cultura de hurto que allí se presenta. Sin embargo, en los países desarrollados como Estados Unidos y el Reino Unido también se presenta esta problemática, aclarando que los porcentajes de pérdidas varían de país en país dependiendo de diferentes factores como el grado de estabilidad política y los niveles de desarrollo económico y social de sus habitantes. El hurto estimado de energía para algunos países en vía de desarrollo es alto, y varía en un rango del 20% al 30%; contrario a los países desarrollados, los cuales alcanzan valores no superiores al 3.5%. En cualquier país, los montos de estas pérdidas tienen un impacto grande, desde el punto de vista económico. Por ejemplo, en los Estados Unidos en el año de 1998, las pérdidas no técnicas costaron a las empresas de distribución entre USD 1.000 millones y USD 10.000 millones, tomando como base utilidades alrededor de USD 280.000 millones [2, 16]. Para el caso de la EEP S.A. ESP estas pérdidas representan aproximadamente $2.600 millones de pesos mensuales.

Es de resaltar que el problema de las pérdidas no técnicas no solo reside en el predio del consumidor, quien realiza el hurto de energía, sino también en las empresas distribuidoras, las cuales no cuentan con planes eficientes de verificación y control de pérdidas. Adicional a esto, las empresas desconocen, no almacenan o no actualizan los registros e información acerca del comportamiento de sus diferentes usuarios y métodos de fraude más sofisticados [3]. En este sentido, más allá de contribuir a la solución del problema de las pérdidas no técnicas, conocer el comportamiento de los usuarios se está convirtiendo en

un aspecto muy importante en el funcionamiento eficiente de los sistemas actuales de distribución.

En lo que respecta a la detección de pérdidas no técnicas, en la literatura se reportan algunas acciones y estudios adelantados por los Organismos Reguladores (OR) en el sector eléctrico en diferentes países, uno de ellos es mencionado por Romero y Vargas quienes destacan políticas regulatorias comunes para el tratamiento de las pérdidas de energía eléctrica. En el documento *Treatment of Losses by Network Operators, Position Conclusions paper* (2009), se observa la utilización de mecanismos de incentivos para su reducción y definición de costos asociados [4].

Estos mismos autores, reportan por ejemplo que en España se estableció la metodología de retribución en el año de la actividad para el período regulatorio, donde se incorporó un esquema simétrico de incentivos y penalizaciones a la tarifa, según el cumplimiento o no de las metas de disminución de las pérdidas de las empresas distribuidoras. De igual forma se concluye que si las empresas dan pérdidas reales menores a las pérdidas objetivo tienen un incremento en la tarifa de hasta un 1%, en caso contrario tienen una retribución hasta ese mismo valor.

Por tanto, es de resaltar que los trabajos efectuados para la detección de fraudes y pérdidas no técnicas en sistemas de distribución, son limitados en su mayoría ya que las empresas de energía utilizan básicamente dos métodos para detectar usuarios fraudulentos, tales como la instalación de medidores electrónicos en los usuarios finales, con el fin de detectar cualquier irregularidad o alteración del dispositivo de medición instalado y la aplicación de modelos estadísticos de estimación, mediante la evaluación técnica y el diseño económico en las redes de distribución instaladas. Sin embargo, estos métodos imponen altos costos operacionales y requieren el uso extensivo de recursos humanos para su aplicación, con el fin de minimizar este tipo de pérdidas en sistemas de distribución de energía eléctrica [23, 27, 28].

Debido a lo anterior, se han realizado diversas investigaciones con el fin de implementar una metodología eficiente para el análisis y la detección de clientes con consumos irregulares basados en minería de datos y métodos de aprendizaje de máquina, con el fin de encontrar patrones anómalos en el consumo de energía mediante métodos basados en la estadística, la distancia, la densidad, el agrupamiento y la desviación de los perfiles de carga, es decir, el patrón de demanda en el consumo de energía para uno o varios clientes en un periodo de tiempo determinado [5, 29].

Estos métodos son la base para diferentes algoritmos de clasificación utilizados por diferentes autores en la detección de pérdidas no técnicas, tales como conjuntos aproximados, máquinas de soporte vectorial, redes bayesianas, arboles de decisión, métodos aumentados y métodos de votación [5, 17, 28, 30]. Estos algoritmos de clasificación han sido aplicados en diferentes países como Reino Unido [31], España [32], Brasil [33] y Colombia [34], entre otros, en los cuales se ha presentado una reducción en el porcentaje de pérdidas no técnicas, validando la importancia de los algoritmos de clasificación utilizando métodos de aprendizaje de máquina, aplicados al problema de pérdidas no técnicas en sistemas de distribución de energía eléctrica.

1.1 Planteamiento del problema

Las pérdidas no técnicas que experimentan las empresas de distribución tienen grandes impactos en diferentes áreas (operacional, comercial, etc.) e influyen en los resultados económicos y financieros de las mismas. Para el caso de la EEP S.A. ESP, se invierte alrededor de $1.800 millones de pesos anuales en revisiones a predios, por lo que se requiere optimizar dichas actividades. La principal falencia es económica, por la gran cantidad de recursos invertidos en revisiones e inspecciones que se deben realizar de algunos predios, y su baja efectividad [3].

Este último impacto es el más crítico, ya que involucra la reducción de beneficios, la escasez de fondos de inversión en la mejora del sistema de potencia, y la necesidad de

implementar medidas para hacer frente a las pérdidas en el sistema de potencia. Los impactos económicos fluyen desde las empresas distribuidoras de energía que están experimentando incrementos en las pérdidas hacia los usuarios registrados en el sistema comercial. En esos casos, los costos de las pérdidas no técnicas son traspasados a los usuarios para cubrirlas dentro de las operaciones de servicios públicos, los cuales son reconocidos vía tarifa. Por lo tanto, la reducción de las pérdidas no técnicas es crítica para los operadores de las redes de distribución, garantiza que los costos tanto del proveedor como de los usuarios se minimicen y se mejore la eficiencia de la red de distribución**.**

Con base en lo anterior se pretende dar respuesta a la pregunta: ¿Qué metodologías son las más adecuadas para garantizar una mayor efectividad en las revisiones proyectadas a campo y si reducen los costos económicos a la EEP S.A. E.S.P por concepto de revisiones en terreno?

La respuesta a la pregunta de investigación es de especial interés para el grupo de investigación y para la EEP S.A. E.S.P ya que permitirá de una forma práctica automatizar el proceso de generación de campañas a terreno con el fin de determinar posibles usuarios con irregularidades en el equipo de medida, garantizar la reducción en el indicador de pérdidas comercial de la compañía y a su vez reducir costos por concepto de revisiones a realizar en campo.

1.2 Objetivo general

Desarrollar las metodologías **Bagging y Adaboost** que permitan el análisis de datos del sistema comercial de la EEP S.A. ESP, con el fin de identificar predios con irregularidades en su equipo de medida, que, a su vez, permitan reducir las pérdidas no técnicas en el sistema de distribución de la compañía.

1.3 Objetivos específicos

- Documentar el estado del arte sobre metodologías para la reducción de pérdidas no técnicas en sistemas de distribución.
- Seleccionar los parámetros más adecuados y caracterizar la base de datos que contiene la información comercial de clientes regulados de la EEP S.A. ESP.
- Seleccionar la metodología más adecuada para identificar los predios que sean escogidos como candidatos de revisión externa, de acuerdo con la base de datos del archivo de facturación del sistema comercial de la EEP S.A. ESP.
- Validar la metodología presentada en un caso de estudio mediante pruebas en terreno, de acuerdo con los resultados obtenidos con las metodologías aplicadas y garantizar efectividad de las actividades marcadas para revisión.

1.4 Propuesta de solución

Este proyecto se divide en seis (6) etapas, con el fin de alcanzar los objetivos propuestos en este documento. Estas etapas están orientadas a realizar un análisis del desempeño de los dos clasificadores (AdaBoost y Bagging), para la detección de pérdidas no técnicas en el sistema de distribución de la EEP S.A. ESP.

En la primera etapa se realiza un análisis bibliográfico para el modelado de métodos de clasificación, con énfasis en detección de pérdidas en sistemas de distribución con base en los consumos asociados a los usuarios en un periodo de tiempo determinado. En la segunda etapa se seleccionan dos métodos de clasificación basados en la minería de datos, además se seleccionan dos métodos diferentes para obtener los datos de entrenamiento (método basado en conocimiento histórico y método basado en la silueta). Luego, en la tercera etapa se implementan los métodos seleccionados en software Matlab tales como Bagging y Adaboost, el cual sirve como herramienta de trabajo para ejecutar los clasificadores y obtener las respuestas de forma ordenada mediante una lista que contiene las matrículas asociadas a los clientes con irregularidad. En la cuarta etapa se realiza una

interfaz que permite seleccionar el archivo que contiene los usuarios por ciclo, además se selecciona el tipo de uso y su estrato para realizar los métodos de clasificación. Luego, en la quinta etapa se realizan las pruebas a los métodos desarrollados bajo distintos escenarios de prueba y se evidencian los problemas encontrados para el estudio de clasificadores en la detección de pérdidas no técnicas en sistemas de distribución. En la etapa final se reportan los resultados obtenidos al aplicar la metodología propuesta para distintos escenarios de prueba, además se presentan las conclusiones y trabajos a futuro.

1.5 Aportes del proyecto

Este proyecto es importante para la EEP S.A. ESP ya que permitirá optimizar el proceso de análisis de campañas de medida directa y medida especial para clientes regulados, toda vez que garantizará una mayor efectividad en el proceso de revisiones y reducción de costos, por concepto de revisiones a terreno in situ, logrando reducir el indicador estratégico de la compañía como lo es el Indicador de pérdidas Comercial.

En este proyecto se desarrolla un aplicativo ejecutable que permite cargar la base de datos de los usuarios por ciclos. También, clasifica la base de datos en usuarios regulares e irregulares con el fin de encontrar pérdidas no técnicas en el sistema de distribución. Este aplicativo permite a través de los métodos implementados, contribuir a identificar posibles clientes con irregularidades en sus equipos de media de predios comerciales, residenciales en los estratos uno (1) al seis (6), e industriales dentro del mercado regulado de la EEP S.A ESP que permita a la Compañía, disminuir su indicador de pérdidas comercial, indicador estratégico, al igual que proyectar una mayor efectividad de las campañas entregadas para ejecución en terreno.

De igual manera, el aplicativo desarrollado en este proyecto puede ser empleado en otras empresas de energía, además que el principio aplicado puede funcionar como análisis de detección de pérdidas en empresas de gas y de acueducto.

1.6 Estructura del documento

Este proyecto está dividido en cinco (5) capítulos. El capítulo inicial contiene la introducción del proyecto, los objetivos, la propuesta de solución y los aportes del proyecto. En el segundo capítulo se encuentran los aspectos teóricos del proyecto, como antecedentes referentes a los métodos de clasificación en minería de datos. El tercer capítulo corresponde a la metodología propuesta, aquí se presenta la adquisición de datos, el pre procesamiento y procesamiento de datos, el aprendizaje supervisado y el aprendizaje no supervisado. En el cuarto capítulo se presenta la aplicación de la metodología propuesta, los escenarios de prueba, consideraciones previas y resultados. En el capítulo cinco se tienen las conclusiones y las indicaciones de trabajos a futuro. Al final se presentan las referencias citadas en este documento y se anexa el código del aplicativo con una breve descripción.

Capítulo 2

Aspectos teóricos

Las pérdidas de energía en los sistemas de distribución son un problema importante que enfrentan las empresas de energía eléctrica. La prestación del servicio de energía eléctrica en Colombia está segmentada en las actividades de generación, transmisión, distribución y comercialización de energía. Cada una de estas actividades tienen como función: 1) la generación corresponde a la producción de energía mediante el uso de diferentes tecnologías (hidráulica, térmica, solar, eólica, entre otros) ; 2) la transmisión se encarga del transporte de energía a altos niveles de tensión en el Sistema de Transmisión Nacional (STN), para Colombia la transmisión se opera a niveles de tensión entre 200 kV y 500 kV; 3) la distribución comprende el transporte de energía hasta los usuarios finales a niveles de tensión inferiores a 220 kV; 4) la comercialización hace referencia a la compra y venta de energía en el Mercado de Energía Mayorista (MEM) y la venta de ésta a los usuarios finales. Es de anotar que la transmisión y la distribución tienen características de monopolio natural, mientras que la generación y la comercialización pueden operar bajo esquemas de competencia [5].

La regulación actual permite que los generadores y transportadores no enfrenten riesgos con relación a las pérdidas de energía eléctrica, ya que al generador se le paga el total de energía entregada al sistema, y al transportador se le reconoce la totalidad de energía transportada independientemente de su facturación y recaudo.

El comercializador, al ser el agente que interactúa con la oferta y la demanda de energía, es el responsable de la gestión de las compras de energía, de la facturación y del recaudo de los pagos efectuados por los usuarios, actividades en las cuales se presenta la mayor parte de las pérdidas no técnicas ocurridas en el sistema.

De otra parte, el OR es el responsable de planear la expansión, las inversiones, la operación y el mantenimiento de los sistemas de distribución, lo cual hace que pueda gestionar las pérdidas técnicas y parte de las pérdidas no técnicas del sistema.

En cuanto a los usuarios finales, se dispuso la separación entre grandes usuarios (usuarios no regulados) y pequeños usuarios (usuarios regulados), estableciendo la libertad de escogencia del prestador del servicio (comercializador). Vale la pena anotar que la mayor parte de las pérdidas no técnicas son ocasionadas por acciones de algunos usuarios, regulados o no regulados, como fraude en los medidores o conexiones ilegales. [5]

Las pérdidas no técnicas derivadas del hurto de energía u otro tipo de manipulaciones que realizan los usuarios crean todo tipo de inconvenientes para las empresas de distribución de energía eléctrica a nivel mundial, máxime en países en desarrollo, tales como suramericanos y africanos. Las pérdidas ocurren por manipulación y/o falla en los medidores de energía, conexiones ilegales, irregularidades en el proceso de facturación [16]. Cuando la energía eléctrica se obtiene de manera irregular de la red secundaria, se refiere a un hurto de energía. En varios casos, cuando las pérdidas totales de energía del sistema son grandes, se vuelve evidente que las pérdidas no técnicas son serias ya que estas representan un porcentaje importante de las pérdidas totales en el sistema [17].

La energía eléctrica entregada debe ser igual a la energía consumida y registrada por los usuarios; sin embargo, esta situación en la realidad es diferente, ya que las pérdidas ocurren como un resultado integral del proceso de transmisión y distribución. Las pérdidas totales en el sistema están dadas por la diferencia entre la energía entregada y la energía vendida [18] como se muestra en la ecuación (2.1):

$$E_{perdida} = E_{entregada} - E_{vendida} \tag{2.1}$$

De acuerdo con Nagi [19], las pérdidas no técnicas son aquellas que ocurren independientemente de las pérdidas técnicas en el sistema. Estas pérdidas son causadas por acciones externas al sistema eléctrico y también por ciertas cargas o condiciones que los cálculos de las pérdidas técnicas fallan en tener en cuenta. Las pérdidas no técnicas se asocian principalmente al hurto de energía en cualquiera de sus diferentes formas;

adicionalmente y en casos poco frecuentes pueden también ser vistas como cargas de clientes que las empresas de distribución no saben que existen y pasan desapercibidas.
Las pérdidas no técnicas son difíciles de medir ya que a menudo estas no son registradas o actualizadas por los operadores del sistema y por tanto no se tiene información de las mismas. Cabe anotar que las pérdidas del sistema son la suma de las pérdidas técnicas y las pérdidas no técnicas. Existen tres fuentes principales que contribuyen a las pérdidas no técnicas [20]:

- Fallas en los equipos de medida.
- Hurto de energía eléctrica.
- Alteración de bases de datos de los sistemas de información (fraude informático).

De otra parte, las pérdidas causadas por averías son esporádicas y poco frecuentes. Entre los factores que más influyen, se tienen: equipos golpeados por descargas atmosféricas, fin en la vida útil del equipo (> 15 años) y descuidos o planes nulos de mantenimiento de los equipos.

En el hurto de energía se concentra en mayor porcentaje las pérdidas no técnicas, este se realiza en redes de baja tensión, y se extiende por ella, presentándose a niveles residenciales, pequeños y medianos comerciales y algunos sectores industriales. Los factores que contribuyen a las actividades referentes a pérdidas no técnicas se pueden caracterizar de la siguiente manera [21, 22]:

- Manipulación de los medidores de energía para que registren consumos menores.
- Uniones ilegales en los bornes de conexión del medidor para que este no registre el nuevo consumo.
- Registros inadecuados o imprecisos por parte de los medidores de energía.
- Líneas Directas.
- Cuentas de consumo eléctrico incorrectas.
- Incumplimiento en el pago de las cuentas de consumo por parte de los clientes.

- Fallas del medidor y equipos asociados.

2.1 Antecedentes

Existen diversas las estrategias y protocolos que a través de los años se han implementado para reducir las pérdidas no técnicas en las empresas de distribución, algunas de estas estrategias se presentan a continuación [23]:

- Reingeniería de procesos para reducir las pérdidas debidas a los procesos administrativos.
- Revisión completa de los usuarios ubicados en barrios o zonas subnormales, siguiendo rutas preestablecidas para evitar pérdidas por conexiones ilegales, fraude y conexiones clandestinas.
- Instalación de medidores en cajas de policarbonato o de doble compartimiento anti hurto con el objetivo de evitar que sean intervenidos.
- Colocación de sello de plástico y guaya de acero, los cuales son difíciles de violar ya que se encuentran numerados, donde consta la sigla de la empresa sobre relieve para permitir un control inmediato sobre sus usuarios.
- Instalación de conductores anti hurto (cable concéntrico) para evitar las conexiones clandestinas. Dicho conductor está compuesto por un alma de aluminio, una capa de aislante plástico que la rodea, una cubierta conductora de cobre que recubre a la primera capa aislante y otra capa aislante de plástico que se encuentra al exterior del conductor.
- Administrativamente se ejecutan planes especiales de facilidad de pago de deudas acumuladas, condonación de deudas, entre otros.
- Asesorar en el uso racional y eficiente de la energía a fin de lograr que el cliente, una vez ingresado, modere sus consumos, evitando así acumular facturas con las posibles consecuencias de suspensiones de servicios.

Sin embargo, en años recientes las técnicas computacionales que se basan en inferencias estadísticas y el denominado aprendizaje de máquina, han adquirido gran importancia en bases de datos relativamente grandes, debido al tamaño y la complejidad para realizar un análisis manual que permita clasificar de manera óptima los datos almacenados en un periodo de tiempo determinado [29].

A continuación, se presentan algunas referencias relacionadas con la necesidad, el desarrollo y la implementación de técnicas computacionales basadas en métodos de aprendizaje de máquina, enfocados a la detección y clasificación de pérdidas no técnicas en sistemas de distribución y sistemas de transmisión de energía eléctrica en diferentes partes del mundo.

En [4] se presentan políticas regulatorias comunes para el tratamiento de las pérdidas de energía eléctrica en España (2009), utilizando mecanismos de incentivos para su reducción, además se definen los costos asociados a dichas pérdidas. Luego, en [1] se propone un modelo de incentivos para la reducción de pérdidas de energía eléctrica en Colombia (2010), basado en penalizaciones y pautas de remuneración para las empresas comercializadoras de energía que incumplan o que por el contrario, se ajusten a las metas de disminución de pérdidas establecidas por los OR, con el fin de disminuir el indicador global de pérdidas técnicas y no técnicas en sistemas de distribución de energía eléctrica. Debido a los modelos de incentivos propuestos, se hace evidente la necesidad de desarrollar e implementar técnicas computacionales que permitan mejorar la eficiencia en la detección de pérdidas no técnicas, de manera que reduzca el indicador global de pérdidas en sistemas de distribución.

En [28] se presenta un método de identificación de fraudes basado en arboles de decisión, por parte de los consumidores de energía en un sistema real de distribución en Brasil (2004). Este artículo presenta una matriz eficiente para la asignación de etiquetas, identificando los usuarios sospechosos mediante el algoritmo C4.5 para la clasificación de los datos, ya que este es un algoritmo que permite utilizar variables continuas tanto en

la etapa de entrenamiento como en la etapa de pruebas para la asignación final de etiquetas a los usuarios conectados al sistema de distribución.

En [24] se propone una metodología que usa históricos de los perfiles de carga de los usuarios para identificar posibles consumos anormales mediante la minería de datos basado en un método de aprendizaje de máquina. Asimismo, se estudia un método basado en algoritmos de clasificación como el Naive Bayesian y el árbol de decisión [25], con el fin de identificar posibles irregularidades que pudieran desencadenar en pérdidas no técnicas.

En [2] se desarrolla e implementa una metodología probabilística para la estimación de pérdidas técnicas y no técnicas en un alimentador en presencia de variaciones de carga para un sistema de distribución con datos reales, mediante la instalación de medidores a lo largo del sistema de distribución en Brasil en el año 2012, de manera que el sistema se divide en varios subsistemas permitiendo una estimación más precisa de las pérdidas en el sistema estudiado. La metodología propuesta en esta referencia realiza una comparación entre el consumo de energía medida en el alimentador con respecto a la energía facturada por la empresa comercializadora más las pérdidas técnicas estimadas en el alimentador y la red eléctrica utilizando datos estadísticos, como la media y la varianza. Luego se realiza el balance de energía y se estiman las pérdidas no técnicas en el sistema. Esta técnica entrega buenos resultados, sin embargo, requiere dispositivos de medición adicionales conectados a lo largo del sistema, lo cual presenta ciertos costos de inversión relacionados con los nuevos dispositivos que se deben conectar a lo largo del sistema de distribución de energía eléctrica.

La referencia [3] presenta una metodología para la detección de fraudes en el consumo de energía, utilizando series en el tiempo independientes de las estaciones del año, para realizar curvas de carga en usuarios conectados a un sistema de distribución en Siberia en el año 2015, con datos históricos que incluyen un grupo de clientes que fueron atrapados robando energía durante el periodo de estudio. Este artículo propone identificar los

clientes con caídas anómalas en la energía consumida (el síntoma más frecuente de una pérdida no técnica en un cliente), es decir, el método detecta cambios considerables que se reflejan en la disminución del consumo de energía por los clientes registrados en la empresa comercializadora, utilizando un control estadístico basado en el consumo histórico de energía por parte de los usuarios conectados al sistema de distribución. La filosofía de este método ya había sido utilizada en España en el año 2011 [5], con una variación en la técnica de detección de clientes con caídas anómalas en la energía consumida, mediante un análisis temporal usando el coeficiente de Pearson, además se usan redes bayesianas y árboles de decisión para detectar otros tipos de patrones de pérdidas no técnicas con clientes reales de la base de datos de Endesa Company. Actualmente, el sistema se encuentra en operación.

A continuación, se presenta una breve descripción de los métodos de clasificación más utilizados en la minería de datos, los cuales han sido explorados por diferentes autores en el problema de pérdidas no técnicas en sistemas de distribución de energía eléctrica y en diferentes áreas de investigación como la clasificación de cuerpos celestes, caracterización de enfermedades, diagnóstico de fallas, predicción de fenómenos naturales, detección de fraudes, entre otros [22].

C4.5: los árboles de decisión generados por este algoritmo pueden ser usados para clasificación, y por lo tanto siempre se refiere a él como un clasificador estadístico. El algoritmo tiene las siguientes características: permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas. Los árboles son menos frondosos, ya que cada hoja cubre una distribución de clases y no una clase en particular. Utiliza el método “divide y vencerás” para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento [7].

Agrupamiento de k vecinos más cercanos (k-means): el algoritmo se inicializa escogiendo puntos iniciales, que se conoce como centroides, los cuales se pueden generar de forma aleatoria o mediante valores fijos preestablecidos. A partir de esto, el algoritmo asigna

cada dato al centroide más cercano, luego calcula de nuevo los centroides de cada grupo y reasigna cada dato al centroide más cercano. Cuando las asignaciones no varían más, el algoritmo converge en un agrupamiento óptimo de vecinos más cercanos [8].
Máquinas de Soporte Vectorial (MSV): las máquinas de soporte vectorial (Support Vector Machines) son un conjunto de algoritmos de aprendizaje no supervisado desarrollados por Vladimir Vapnik [9]. Estos conjuntos de algoritmos están relacionados propiamente con problemas de clasificación y regresión. Dado un conjunto de datos de entrenamiento, se etiquetan las clases y se entrena el algoritmo para construir un modelo que prediga la clase de un nuevo dato.

Algoritmo a priori: este es un algoritmo que se usa para encontrar reglas de asociación en un conjunto de datos. Este algoritmo se basa en el conocimiento previo de los conjuntos frecuentes, esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia [10]. El algoritmo opera en bases de datos que contienen transacciones (por ejemplo, recopilación de objetos adquiridos por consumidores o detalles sobre la frecuencia de acceso a páginas web). Dado un umbral, el algoritmo a priori identifica los conjuntos de elementos que son subconjuntos de al menos C transacciones en la base de datos

Algoritmo de Maximización de la Esperanza (ME): este algoritmo se usa para encontrar estimadores de máxima verosimilitud en parámetros con modelos probabilísticos que dependen de variables no observables. El algoritmo ME alterna pasos de esperanza (paso E), donde se computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y un paso de maximización (paso M), donde se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E. Los parámetros que se encuentran en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite, hasta llegar a una convergencia pre establecida. [11].

Algoritmo PageRank: este algoritmo es una marca registrada y patentada por Google el 9 de enero de 1999 que ampara una familia de algoritmos utilizados para asignar de forma

numérica la relevancia de los documentos (o páginas web) indexados por un motor de búsqueda. Sus propiedades son muy discutidas por los expertos en optimización de motores de búsqueda. El sistema PageRank es utilizado por el popular motor de búsqueda Google para ayudarle a determinar la importancia o relevancia de una página. Fue desarrollado por los fundadores de Google, Larry Page (apellido, del cual, recibe el nombre este algoritmo) y Sergey Brin, en la Universidad de Stanford [12].

Clasificador Bayes Naive: el clasificador Bayes Naive, es un algoritmo de clasificación supervisado. El clasificador se basa en el teorema de Bayes. Este método presenta un análisis cualitativo y cuantitativo de la instancia clasificada. En el aspecto cualitativo porque relaciona los atributos ya sea en forma casual o señalando la correlación que existen las diferentes variables y en el cuantitativo porque da una medida probabilística de la importancia de cada variable en el problema [13]. Otros algoritmos como árboles de decisión o redes neuronales no ofrecen una medida cuantitativa de la clasificación.

$$p(\ C\ |F1, \ldots . ; Fn) = \frac{p(C)p(F1,\ldots.,Fn\ |C)}{p(F1,\ldots.,Fn)} \tag{2.2}$$

El clasificador Bayes Naive presenta las siguientes características: cada variable observada modifica la probabilidad de hipótesis de cada clase, es robusto al ruido que pueda presentarse en la etapa de entrenamiento, requiere poca cantidad de datos de entrenamiento para estimar los parámetros necesarios y es rápido en comparación a métodos más sofisticados.

Redes Neuronales Artificiales: las Redes Neuronales Artificiales (RNA) han sido desarrolladas como generalizaciones de modelos matemáticos y computacionales que se basan en las redes neuronales biológicas. Una RNA está conformada por múltiples unidades llamadas Neuronas Artificiales (NA), las cuales se interconectan y operan en paralelo para procesar información. Este proceso se realiza bajo las siguientes suposiciones [14]: el procesamiento de la información ocurre en las neuronas artificiales,

las señales pasan entre las neuronas por medio de enlaces de conexión, cada enlace de conexión tiene asociado un peso el cual multiplica la señal transmitida. Cada neurona aplica una función de activación (usualmente no lineal) para determinar la señal de salida.

Algunas de las metodologías más atractivas y que se acomodan al caso de estudio (determinación de pérdidas no técnicas en sistemas de distribución) son las metodologías Bagging y Adaboost. La metodología Bagging adopta la distribución Bootstrap. Esta distribución se define como un re muestreo propuesto por Bradley Efron en 1979 el cual se utiliza para aproximar la distribución en el muestreo de un estadístico, con el fin de generar diferentes modelos base y las estrategias de agregar las salidas a los algoritmos base, que se interpreta como "elegir", en el caso de clasificación y "promediar", en el caso de regresión. La idea básica es re muestrear los datos y calcular las predicciones sobre el conjunto de datos re muestreados. Este método se basa en árbol de decisiones [15]. Por otra parte, la metodología Adaboost, construye un modelo fuerte a partir de una combinación lineal de modelos base o modelos débiles. Esta metodología es considerada un clasificador suave que se vuelve robusto, ya que se ejecuta varías veces dentro de su rutina [26].

2.2 Minería de datos

El proceso de minería de datos se define como el conjunto de técnicas que permiten examinar grandes bases de datos, ya sea de forma semiautomática o automática, con el objetivo de encontrar patrones, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Las fases más importantes en minería de datos son [22]:

- Clasificación: Modelo que permite obtener datos de clase desconocida a clase concreta.

- Regresión: Modelo que permite predecir el valor numérico de alguna variable.
- Agrupamiento: Asociación de datos bajo un criterio determinado.

El proceso de minería de datos involucra la integración de técnicas de múltiples disciplinas tales como tecnologías de bases de datos, estadística, aprendizaje de máquina, reconocimiento de patrones, visualización de datos, recuperación de la información, procesamiento de imágenes y análisis espacial o temporal de los datos [19].

2.2.1 Algoritmo Bagging

El nombre Bagging viene de la abreviación de Bootstrap aggregation [21]. Este método se compone de dos factores claves, Bootstrap y Aggregation.

El primero, es un método de re muestreo propuesto por Bradley Efron en 1979, el cual se utiliza para aproximar la distribución en el muestreo de un estadístico. Se usa frecuentemente para aproximar la varianza de un análisis estadístico, así como para construir intervalos de confianza o realizar contrastes de hipótesis sobre parámetros de interés. El segundo componente es un proceso de recopilación, en el cual las instancias de la experiencia son agrupadas en un conjunto.
El algoritmo Bagging adopta la distribución Bootstrap para generar diferentes modelos base y las estrategias de agregar las salidas a los algoritmos base, lo cual se interpreta como votar en el caso de clasificación y promediar en el caso de regresión.

El método Bootstrap realiza una distribución de probabilidad empírica a partir de la muestra, y asigna una probabilidad de $1/m$ a cada punto de la muestra. Luego, extrae una muestra aleatoria denominada “re muestra” b y se calcula el estadístico de interés mediante la ecuación (2.2) para la muestra b, el cual representa la media de los valores estadísticos calculados en las B re muestras.

$$\hat{\theta}^{*}_{(b)} = \frac{\sum_{b-1}^{B} \hat{\theta}^{*}_{b}}{B} \tag{2.3}$$

Para predecir una instancia, tomando como ejemplo la clasificación, se recolectan las salidas de los *B* modelos base, vota y elije la etiqueta ganadora como la predicción. En la Figura 1, X representa los datos de entrenamiento, los cuales alimentan los diferentes modelos base $Y_M(x)$, que generan las salidas $T_M(x)$. El resultado final $Y_M(x)$, depende del voto mayoritario de cada una de las salidas de los modelos base.

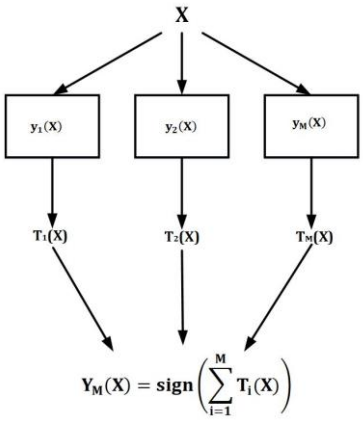


Fig.1. Esquema de la estructura del algoritmo Bagging

Bagging puede manejar tanto clasificaciones binarias como clasificaciones multi-clases. La idea básica es re muestrear los datos y calcular las predicciones sobre el conjunto de datos re muestreados. Al promediar varios modelos, conjuntamente se obtiene un mejor ajuste debido a que se mitigan tanto los modelos con sesgo como los modelos con alta varianza. La agregación de las predicciones de múltiples clasificadores con el objetivo de mejorar la precisión se denomina *ensemble methods*. Si el algoritmo predice datos categóricos, entonces el voto de la mayoría dará la clase dominante o con mejor predicción. Si se está realizando predicción sobre datos numéricos, entonces se realizará la media sobre las predicciones [21]. A continuación, se presenta la estructura general de método Baggiong, el cual se basad en árboles de decisión, donde $\hat{\theta}^{*}_{b}$ es la distribución Bootstrap.

Estructura del clasificador Bagging
Entrada: $conjunto\ de\ datos\ depurados\ \theta = \{X, Y\}$ ***para todo*** $b = 1{:}\ B$ ***hacer*** $h_b = l\left(\theta, \hat{\theta}_b^*\right)$ *, entrena el clasificador* ***fin para***
Salida: $Y_M(X) = argmax \sum_{b=1}^{B} i^2 = \mathbb{I}(h_b(x) = y)$

2.2.2 Algoritmo Adaboost

AdaBoost es la forma corta de las palabras Adaptive Boosting. Este algoritmo fue desarrollado por Freund & Schapire (1995) y se considera como el algoritmo más influyente en el caso de los métodos aumentados. Adaboost construye un modelo fuerte a partir de una combinación lineal de modelos base o modelos débiles. La premisa básica es que múltiples modelos o clasificadores débiles pueden combinarse para generar un modelo conjunto más preciso, que se conoce como modelo fuerte, aun si los modelos débiles tienen un desempeño algo mejor de forma aleatoria.

En la Figura 2, X representa los datos de entrenamiento, cada modelo base Y_M (x), es entrenado con un valor ponderado que depende del conjunto de entrenamiento, los pesos Wn (x) dependen del desempeño del modelo base anterior. Una vez que todos los modelos han sido entrenados, estos se combinan para obtener el modelo final (clasificador final) Y_M (x). Adaboost es un clasificador suave que se vuelve robusto ya que se ejecuta varias veces dentro de su rutina.

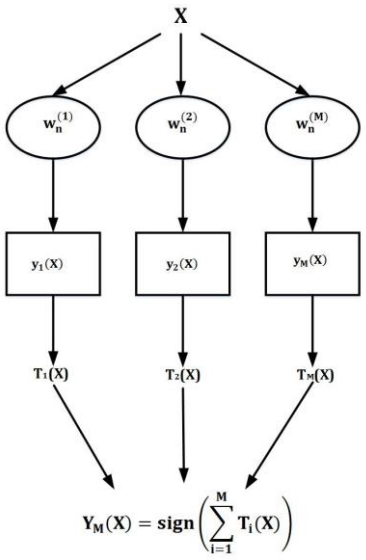


Fig.2. Esquema de la estructura del algoritmo Adaboost

A continuación, se presenta la estructura general de método Adaboost, donde Z_b es una normalización para la distribución [6].

Estructura del clasificador Adaboost
Entrada: *conjunto de datos depurados* $\theta = \{X, Y\}$ $\hat{\theta}_1(x) = 1/m$ *, inician los pesos de la distribución* ***para todo*** $b = 1{:}\,B$ ***hacer*** $h_b = l(\theta, \hat{\theta}_b)$ *, entrena el clasificador* h_b $e_b = P_{(x)\sim\hat{\theta}_b}(h_b(x) \neq f(x))$ *, evalúa el error de* h_b ***si*** $e_b > 0.5$ ***entonces*** $\alpha_b = \frac{1}{2}\log\left(\frac{1-e_b}{e_b}\right)$ *, determina el peso para* h_b $\hat{\theta}_{b+1}(x) = \frac{\hat{\theta}_b(x)exp^{(-\alpha_b f(x)h_b(x))}}{Z_b}$ *,actualiza la distribución* ***fin si*** ***fin para***
Salida: $Y_M(X) = sign\sum_{b=1}^{B}\alpha_b h_b(x)$

Capítulo 3

Metodología propuesta

La metodología propuesta se divide en tres (3) etapas como se ilustra en la figura 3. De acuerdo con los resultados obtenidos se aplicará la metodología en el análisis de campañas en la EEP S.A. ESP, de modo que permita la optimización de recursos y de costos en los procesos de revisiones, y a su vez, la reducción del indicador estratégico de pérdidas de la EEP S.A. ESP.

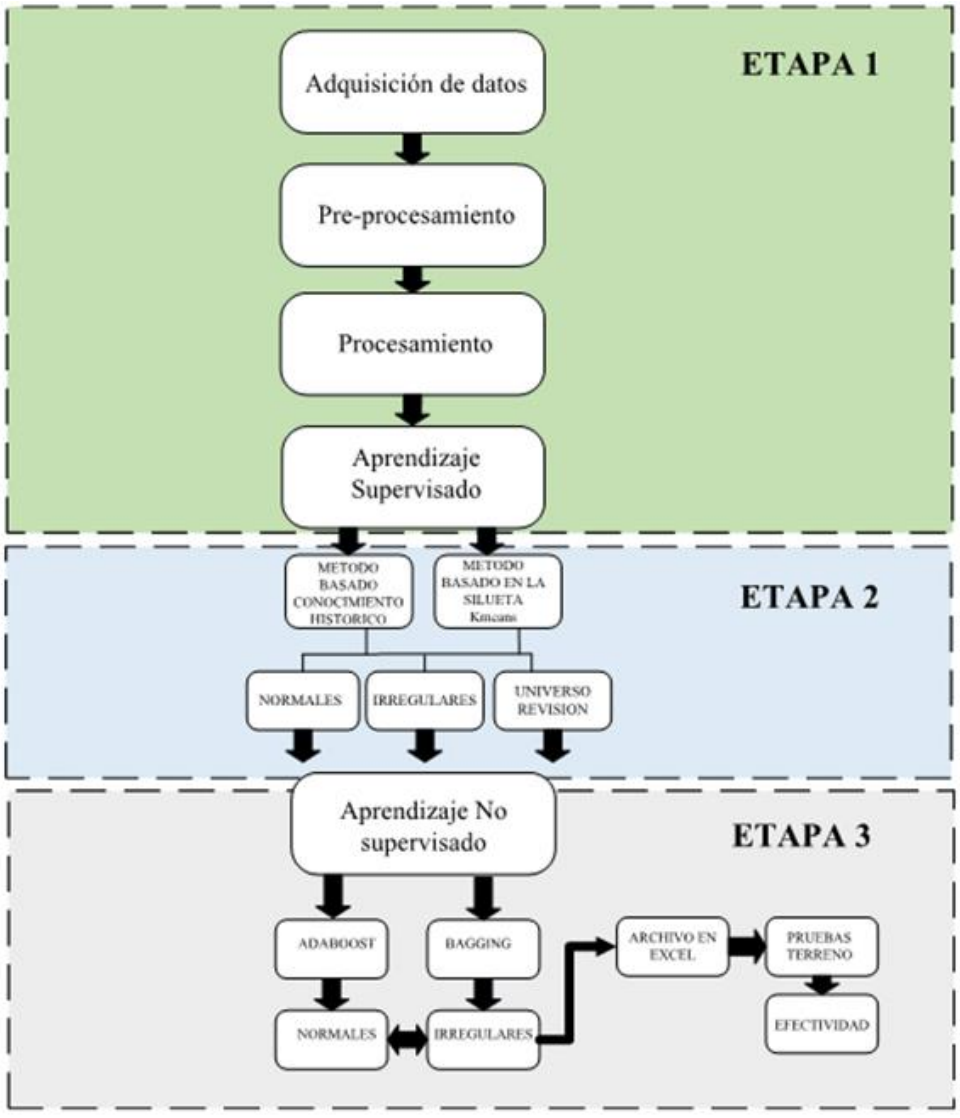


Fig. 3. Metodología propuesta

3.1 Adquisición de datos

Se ingresan datos reales con información de facturación del sistema comercial de la EEP S.A. ESP, con el fin de realizar pruebas piloto que permitan detectar si con la metodología aplicada se garantiza una efectividad positiva que conlleve a la detección de irregularidades en el sistema de distribución de la compañía.

Las etapas de pre-procesamiento y procesamiento representan el núcleo central de la metodología que se propone, ya que en la primera etapa se realiza un análisis estadístico a los datos que se usan como insumo, con el fin de depurarlos y obtener mejores resultados en la etapa posterior de procesamiento. Por lo demás, se analiza la base de datos del sistema comercial de la compañía y se incluyen los datos relevantes objeto del estudio por un período de un año, ya que la EEP S.A. ESP lleva un control estricto de los consumos de los usuarios con los diferentes tipos de lectura y asignación manual de etiquetas (regulares e irregulares) tomadas en campo, sin embargo estos datos no eran usados de manera óptima por la EEP S.A. ESP para la detección de pérdidas negras debidas a irregularidades y/o robos por parte de los usuarios de la empresa. En adición, los resultados obtenidos en esta investigación demuestran que un periodo de 12 meses es suficiente para realizar el análisis estadístico y presentar la asignación de etiquetas para los clientes de la empresa mediante la aplicación desarrollada en este proyecto.

3.2 Preprocesamiento de datos

En esta etapa se accede a la base de datos de la EEP S.A. ESP, con el fin de obtener los datos más relevantes, e integrarlos en un archivo. mat, el cual constituye el insumo principal de las etapas posteriores. El archivo. mat contiene tres tipos de usuarios que son: normales de entrenamiento, irregulares de entrenamiento y pendientes por definir (candidatos para revisión). Los usuarios normales e irregulares de entrenamiento son los usuarios conocidos a priori por la EEP S.A. ESP, los cuales se han etiquetado luego de realizar una revisión estricta de los predios seleccionados. Los usuarios pendientes por

definir son los usuarios que ingresan a los clasificadores, con el fin de etiquetarlos y generar un archivo en Excel que contenga la identificación de los usuarios irregulares.
La base de datos contiene los consumos eléctricos en cada periodo (kWh) y observación de lectura, y otras variables representativas (ubicación, tipo de usuario, actividad económica, etc.), como se muestra en la Tabla 1:

Tabla. 1. Formato consolidado de los consumos eléctricos para n periodos.

MATRICULA	VARIABLES REPRESENTATIVAS			CONSUMOS DE ENERGIA EN kWh y SOL. CONSUMO			
	UBICACIÓN	TIPO	ESTRATO	Consumo mes 1	Sol. Consumo 1	… Consumo n	Sol consumo n
123456	Urbano	Comercial	1				
348753	Rural	Residencial	2				
		Industrial	3				
			4				
			5				
			6				

3.3 Procesamiento de datos

En esta etapa, se analiza y depura la información de entrada con el fin de separar en grupos los diferentes tipos de usuarios como residenciales en los diferentes estratos, industriales y comerciales. Luego se aplican filtros para remover datos atípicos que puedan estar presentes, y así garantizar que la información resultante sea lo más limpia posible. Los filtros que se usan para la etapa de procesamiento se dividen en 3 partes:

- Se reemplazan los datos faltantes o espacios vacíos en los consumos de la base de datos, por consumos iguales a cero.
- Se remueven los usuarios que tienen la mayoría de los consumos en cero.
- Se remueven los datos atípicos – Estos se incluyen para revisión.

3.4 Aprendizaje supervisado

En esta etapa se separan los usuarios de la base de entrenamiento en dos grupos, grupo de usuarios irregulares y grupo de usuarios normales. Los grupos se seleccionan mediante las dos metodologías propuestas que se describen a continuación.

El primer método (Método Basado en conocimiento histórico), utiliza el conocimiento a priori de los usuarios de entrenamiento separado en dos grupos, regulares e irregulares. Esta base de datos es entregada por la EEP S.A. ESP y contiene el registro de los usuarios, clasificados luego de realizar una revisión en el predio de varios clientes, con el fin de obtener una clasificación real que sirva como base de entrenamiento para los métodos de aprendizaje supervisado. Los datos de prueba corresponden a los registros de los usuarios que no han sido clasificados por EEP S.A. ESP.

El segundo método (Método basado en la Silueta), utiliza un algoritmo de agrupamiento que divide los datos de entrenamiento en un número n de grupos con su correspondiente centroide por grupo, donde n es determinado por el criterio de la silueta buscando el número óptimo de grupos en que se debe dividir los datos de entrenamiento. Luego de dividir en n grupos, se clasifica el grupo más grande como usuarios normales, mientras que el grupo menor se clasifica como usuarios irregulares, de manera que los $n - 2$ grupos restantes que quedan fuera de la clasificación inicial corresponden a los datos de prueba, los cuales serán objeto de estudio y posible revisión.

Para el caso concreto el criterio de agrupamiento se divide la base de datos en $n=6$ grupos (este criterio separa la base de datos objeto del estudio en grupos, los cuales presentan características y tendencias de consumos similares entre ellos, empezando por $n = 3$ hasta llegar a $n = 6$ donde converge el método de agrupamiento), parámetro definido en el software. Este criterio de agrupamiento se basa en el centroide de los grupos previamente definidos para etiquetar y clasificar los datos en cada uno de los grupos.

3.5 Aprendizaje no supervisado

Los métodos de clasificación deben construir un modelo con los datos de entrenamiento, para evaluarlo en los datos de prueba y predecir a que grupo pertenece cada usuario de la base de datos de prueba. Cada método de clasificación tiene dos etapas, las cuales corresponden al entrenamiento y la clasificación.

La etapa de entrenamiento crea un modelo que aprende a dividir los usuarios en dos categorías (normales e irregulares). Es muy importante que los datos de entrenamiento se encuentren bien depurados, debido a la sensibilidad de los métodos ante datos espurios. Luego se tiene la etapa de clasificación, en la cual se prueban los modelos creados, con el fin de obtener las etiquetas para cada usuario en la base de datos de prueba.

Para la metodología presentada se utilizan dos algoritmos de clasificación, los cuales en orden de aplicación son: Bagging y AdaBoost.

3.6 Medida de desempeño

En esta última etapa, se validan los resultados arrojados por cada algoritmo y se genera una lista final de usuarios en Excel, el cual se identifica por matrícula y representan los predios con posibilidad de hurto de energía o daño en su equipo de medida. Se usa la eficiencia como índice de desempeño, como se muestra a continuación:

$$eficiencia = \frac{I}{U} * 100\% \tag{3.1}$$

donde,

I = # usuarios con irregularidad

U = # usuarios candidatos como irregulares

Otra manera de medir la eficiencia es utilizar como denominador el número total de los candidatos como irregulares obtenidos en el aprendizaje no supervisado más el número total de usuarios encontrado como irregulares en el aprendizaje supervisado (criterio de la silueta). Sin embargo, esta medida de desempeño no fue utilizada debido a que como

trabajo futuro se pretende manejar una base de entrenamiento fija, que no dependa de la base de datos de la compañía ni del criterio de la silueta, sino que tenga un clasificador previamente entrenado para etiquetar cualquier base de datos, no solo de empresas distribuidoras de energía sino también para agua y gas ya que los comportamientos o tendencias de consumo en estas empresas son similares para usuarios irregulares.

$$eficiencia = \frac{I}{U + US} * 100\% \tag{3.2}$$

donde,

US = # usuarios candidatos como irregulares

Capítulo 4

Aplicación de la metodología propuesta

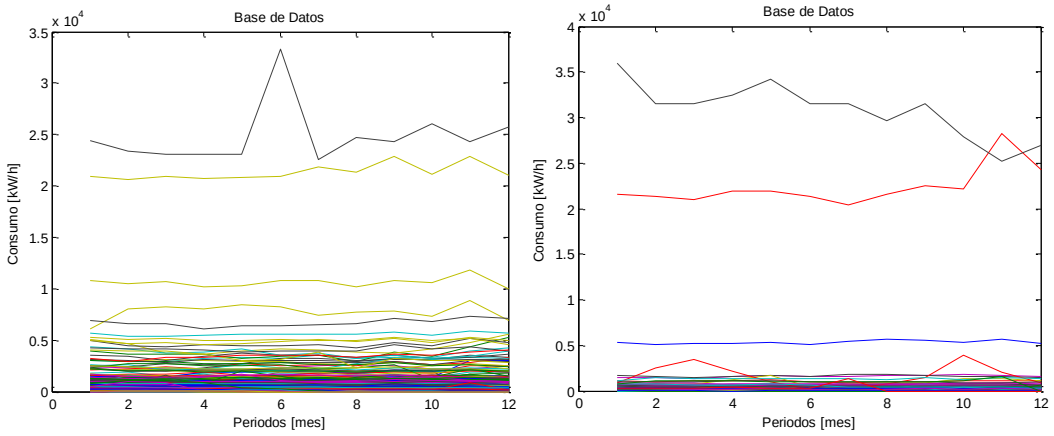
La metodología aplicada usa la base de datos de la EEP S.A. ESP, para encontrar los usuarios irregulares aplicando minería de datos, y evaluando la eficiencia de los métodos de clasificación.

A continuación, se presentan dos escenarios de prueba correspondientes al periodo de un año entre el 2016 y el 2017 con diferentes ciclos (corresponde a la nomenclatura entregada desde el sistema comercial a los usuarios sectorizados por ubicación geográfica en el municipio de Pereira) y tipos de usuario. El primer escenario de prueba se presenta el ciclo 10 con usuarios comerciales, este escenario contiene 552 usuarios antes de realizar la etapa de procesamiento. El segundo escenario de prueba se realiza para el ciclo 16 con usuarios comerciales, los cuales corresponden a 87 usuarios antes de realizar la etapa de procesamiento.

4.1 Base de datos

Primera prueba – Ciclo 10: En la figura 4a se muestra la base de datos del ciclo 10 con 8947 usuarios, de los cuales 552 usuarios corresponden a comerciales. Se escogen estos usuarios ya que de acuerdo con la experiencia y el comportamiento respecto a recuperación de energía presentan una probabilidad mayor a los usuarios residenciales en cuanto a la ejecución de irregularidades o fraudes en el equipo de medida. Esta figura muestra la información de la base de datos sin datos atípicos y sin quitar ceros, es decir, sin filtros.

Segunda prueba – Ciclo 16: En la figura 4b se muestra la base de datos del ciclo 16 con 5519 usuarios, de los cuales 87 usuarios corresponden a comerciales. Esta figura muestra la información de la base de datos sin datos atípicos y sin quitar ceros.



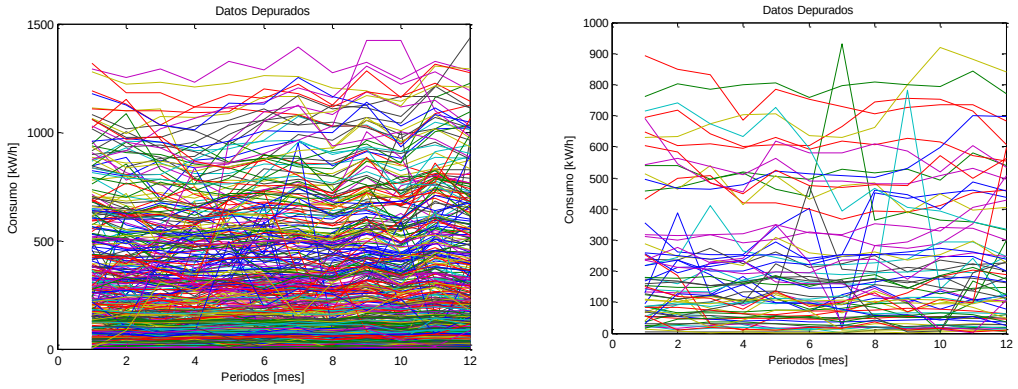
(a) (b)

Fig. 4. Base de datos para los usuarios comerciales del ciclo 10 (a) y ciclo 16 (b)

4.2 Filtros aplicados al procesamiento

Se aplican los filtros mencionados en la metodología, con el fin de eliminar datos atípicos en la base de datos.

Primera prueba – Ciclo 10: En la figura 5a, se muestra la base de datos luego de realizar los filtros para depurar y eliminar los usuarios atípicos. Luego de realizar la etapa de filtrado, se tienen 395 usuarios para iniciar el aprendizaje supervisado.



(a) (b)

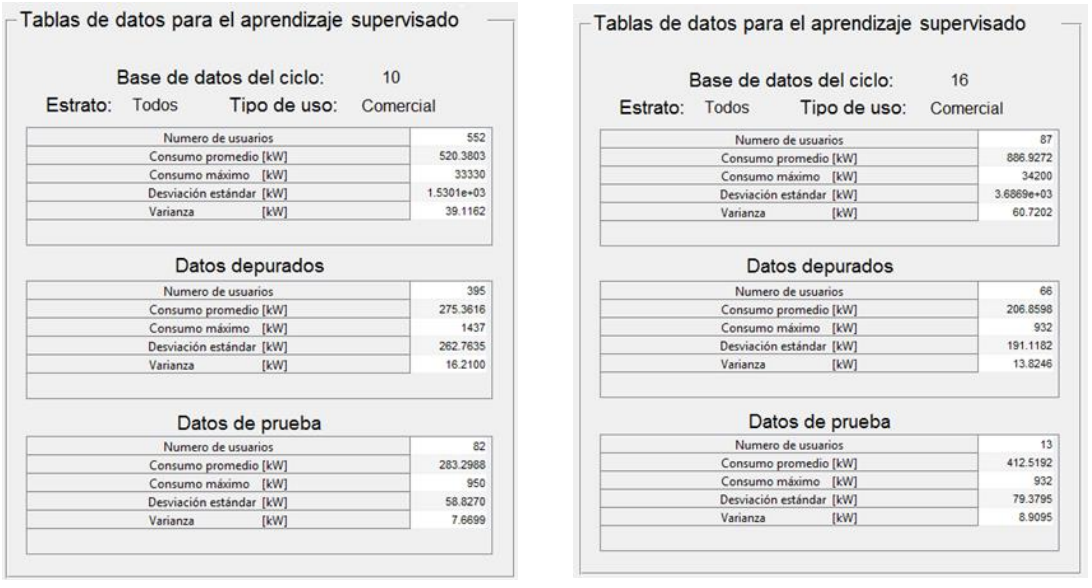
Fig. 5. Datos de depurados para los usuarios comerciales del ciclo 10 (a) y ciclo 16 (b)

Segunda prueba – Ciclo 16: En la figura 5b, se muestra la base de datos luego de realizar los filtros para depurar y eliminar los usuarios atípicos. Luego de realizar la etapa de filtrado, se tienen 66 usuarios para iniciar el aprendizaje supervisado.

4.3 Resultados del aprendizaje supervisado

Se proponen dos métodos diferentes para el aprendizaje supervisado. El primer método de agrupamiento utiliza el conocimiento de a priori del comportamiento de usuarios catalogados como normales e irregulares, mediante pruebas de campo. El segundo método de agrupamiento separa en varios grupos la base de entrenamiento y selecciona el mayor como el grupo normal, mientras el menor grupo queda etiquetado como irregular.

En este proyecto, solo se presentan los resultados obtenidos con la segunda metodología, debido a su alta eficiencia comparada con la primera metodología de aprendizaje no supervisado, ya que utiliza el agrupamiento de k vecinos más cercanos (k-means) y el criterio de la silueta para encontrar los usuarios irregulares. En las figuras 6a y 6b, se presentan los resultados estadísticos para el aprendizaje supervisado con la segunda metodología propuesta, en usuarios comerciales para el ciclo 10 y el ciclo 16 respectivamente.



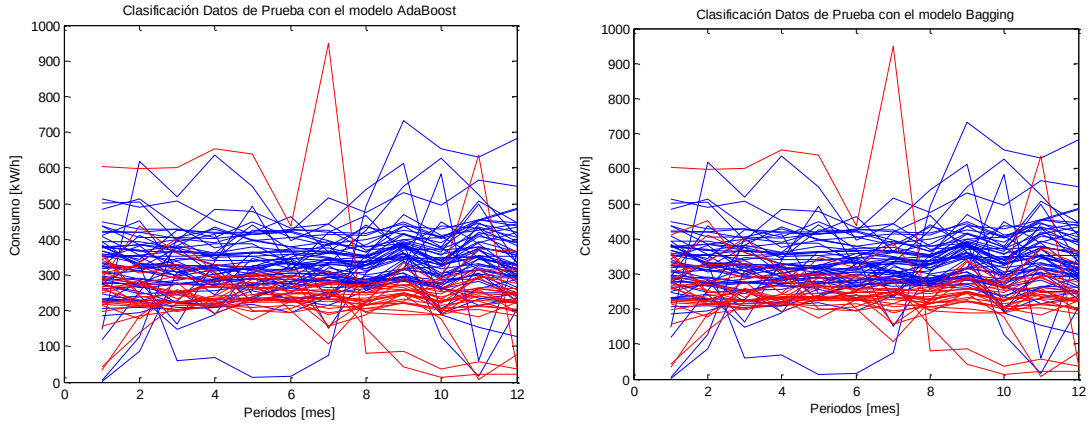
(a) (b)

Fig. 6. Resultados del aprendizaje supervisado para el ciclo 10 (a) y para el ciclo 16 (b)

4.4 Resultados del aprendizaje no supervisado

A continuación, se presenta la base de datos de entrenamiento (Base de datos general luego de aplicarse los filtros de consumos ceros y consumos atípicos del sistema comercial, éste resultado se ingresa al proceso de asignación de etiquetas para seleccionar dicha base de datos) y de prueba respecto a cada escenario, con la segunda metodología propuesta para el aprendizaje supervisado (Método basado en la silueta).

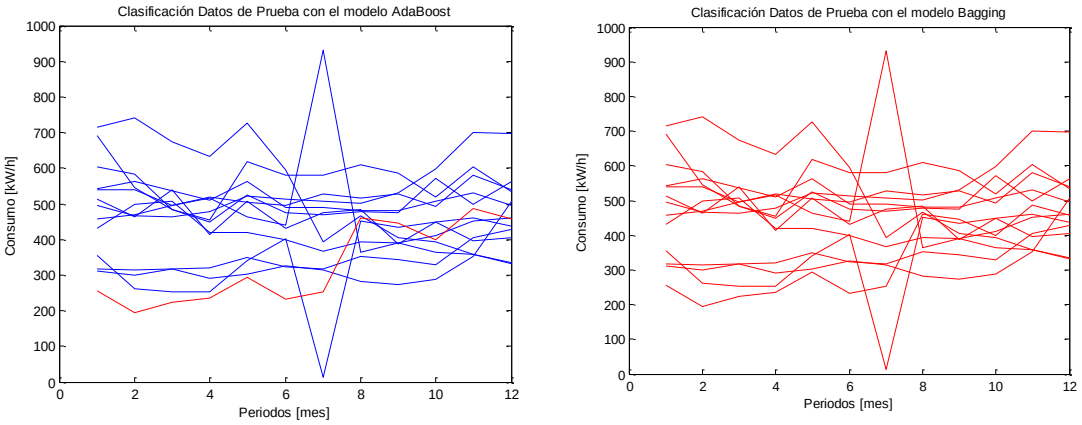
Se ingresan los datos depurados de entrenamiento y de prueba en los métodos de clasificación, para etiquetar los usuarios con comportamiento irregular de acuerdo con las metodologías propuestas. En las figuras 7a y 7b se muestran los resultados de la clasificación con el algoritmo AdaBoost y Bagging respectivamente, para el ciclo 10 con 82 usuarios tipo comerciales.



(a) (b)

Fig. 7. Resultados con AdaBoost ciclo 10 (a) y resultados Bagging ciclo 10 (b)

En las figuras 8a y 8b, se presentan los resultados de la clasificación con el algoritmo AdaBoost y Bagging respectivamente, aplicados a 13 usuarios tipo comerciales en el ciclo 16.



(a) (b)

Fig. 8. Resultados con AdaBoost, ciclo 16 (a) y resultados con Bagging, ciclo 16 (b)

Las figuras; 9, 10, 11 y 12, presentan el resumen estadístico de los resultados de clasificación con los algoritmos AdaBoost y Bagging, aplicados a usuarios tipo comerciales en el ciclo 10 y el ciclo 16 de la Empresa de Energía de Pereira.

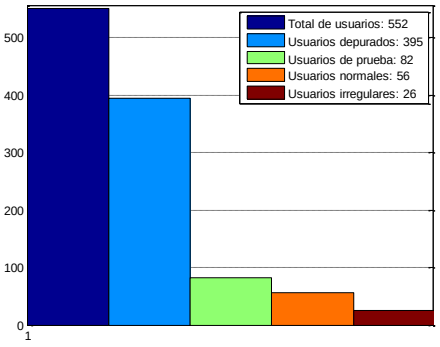


Fig. 9. Resultados Adaboost para el ciclo 10

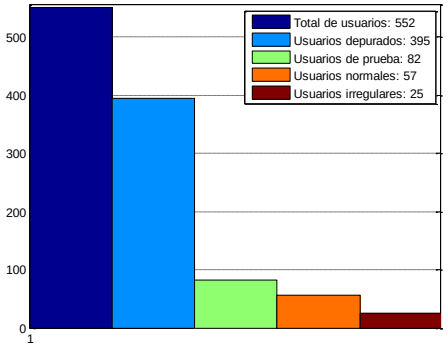


Fig. 10. Resultados Bagging para el ciclo 10

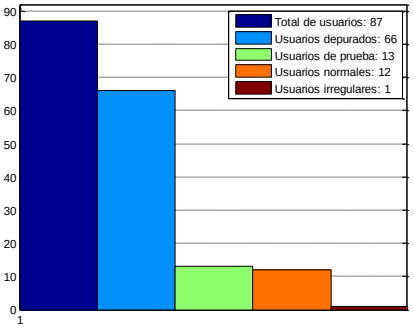


Fig. 11. Resultados Adaboost para el ciclo 16

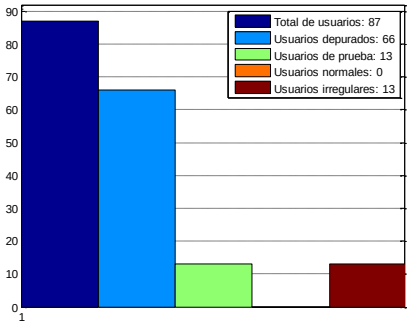


Fig. 12. Resultados Bagging para el ciclo 16

4.5 Medida de desempeño

La aplicación creada en esta tesis entrega una hoja de Excel con las matrículas de los clientes irregulares. La medida de desempeño del método se mide realizando las revisiones en campo con la base de datos que contiene los clientes irregulares entregados por la aplicación. En la tabla 2, se presenta el resumen de los datos encontrados mediante el desarrollo de la metodología propuesta.

Tabla 2. Resumen de resultados en la aplicación del método propuesto

Escenarios de prueba	Ciclo 10 - Comercial	Ciclo 16- Comercial
Número de usuarios para la base de datos sin depurar	8947	5519
Número de usuarios para la base de datos depurada	552	87
Número de usuarios para la base de entrenamiento	448	76
Número de usuarios para la base de prueba	82	13
Usuarios irregulares encontrados	Adaboost	
	26	3
	Bagging	
	25	13
Usuarios irregulares encontrados por la Empresa	6	1
Eficiencia del método	Adaboost	
	23,08%	33,33%
	Bagging	
	24,00%	7,69%

En las tablas 3 y 4 se puede observar el total de actividades ejecutadas en terreno donde se obtuvo para el ciclo 10 una efectividad del 23.08%, la cual fue producto de 6 revisiones efectivas que representaron $13.1 millones de pesos y 26.51 MWh para la compañía por procesos administrativos de recuperación de energía (PARE). De igual manera una efectividad en el ciclo 16 del 33.33%, garantizando $ 2.8 millones de pesos y 5.7 MWh de energía.

Tabla 3**.** Resultados obtenidos de pruebas en terreno – Ciclo 10

Matrícula	Fecha Revision	Acta	Resultado	Efectiva	Recuperación de energía estimada (kWh)	Energía estimada en Pesos ($)	Costo Revisiones
476895	30/05/2017	601892	Corrección	SI	18.547	$ 9.192.132,00	$ 43.149,68
477125	31/05/2017	601897	Corrección por sellos	NO	0	$ -	$ 43.149,68
1413145	31/05/2017	601898	Cambio Medidor	SI	2192	$ 1.083.383,00	$ 46.629,68
1678382	31/05/2017	601896	Normal	NO	0	$ -	$ 43.149,68
2006877	31/05/2017	2006877	Cambio Medidor	SI	2089	$ 1.035.335,00	$ 46.629,68
480491	01/06/2017	602560	Normal	NO	0	$ -	$ 43.149,68
480541	01/06/2017	602557	Normal	NO	0	$ -	$ 43.149,68
540161	01/06/2017	602554	Cambio Medidor	SI	0	$ -	$ 46.629,68
1243575	01/06/2017	602559	Corrección	SI	3686	$ 1.826.829,00	$ 43.149,68
1423854	01/06/2017	601898	Cambio Medidor	SI	0	$ -	$ 46.629,68
1719640	01/06/2017	602556	Actualización	NO	0	$ -	$ 43.149,68
480814	02/06/2017	602564	Normal	NO	0	$ -	$ 43.149,68
483602	02/06/2017	602551	Antitécnica	NO	0	$ -	$ 43.149,68
486175	02/06/2017	602552	Corrección por sellos	NO	0	$ -	$ 43.149,68
486597	02/06/2017	602566	Corrección por sellos	NO	0	$ -	$ 43.149,68
556134	02/06/2017	601890	Normal	NO	0	$ -	$ 43.149,68
694349	02/06/2017	602558	Normal	NO	0	$ -	$ 43.149,68
934364	02/06/2017	601899	Corrección por sellos	NO	0	$ -	$ 43.149,68
941112	02/06/2017	602565	Normal	NO	0	$ -	$ 43.149,68
975052	02/06/2017	601893	Antitécnica	NO	0	$ -	$ 43.149,68
989988	02/06/2017	602555	Antitécnica	NO	0	$ -	$ 43.149,68
1064229	02/06/2017	601895	Corrección por sellos	NO	0	$ -	$ 43.149,68
1109180	02/06/2017	601891	Normal	NO	0	$ -	$ 43.149,68
1173467	02/06/2017	601900	Corrección por sellos	NO	0	$ -	$ 43.149,68
1505817	02/06/2017	602563	Inspección	NO	0	$ -	$ 28.537,16
569012	11/06/2017	604149	Cambio por equipo de mayor precisión	NO	0	$ -	$ 46.629,68
					26.514	$ 13.137.679,00	$ 1.124.679,16

EFECTIVIDAD	**# TOTAL REVISIONES EFECTIVAS *100%** / **# TOTAL REVISIONES**	6 / 26	**23,08%**

Tabla 4. Resultados obtenidos de pruebas en terreno – Ciclo 16

Matrícula	Fecha Revision	Acta	Resultado	Efectiva	Recuperación de energía estimada (kWh)	Energía estimada en Pesos ($)	Costo Revisiones
508473	02/06/2017	602567	11- Antitécnica	NO	0	$ -	$ 43.149,68
571778	02/06/2017	602567	11- Antitécnica	NO	0	$ -	$ 43.149,68
1226331	02/06/2017	602569	2- Cambio de Medidor	SI	5700	$ 2.826.119,84	$ 43.149,68
					5.700	$ 2.826.119,84	$ 129.449,04

EFECTIVIDAD	# TOTAL REVISIONES EFECTIVAS *100% / # TOTAL REVISIONES	1 / 3	**33,33%**

Capítulo 5

Conclusiones y trabajos a futuro

5.1 Conclusiones

Durante el desarrollo de esta aplicación, se pudo evidenciar que de acuerdo con los clientes industriales Regulados y No Regulados, no llevan una curva característica, y al intentar realizar un análisis de consumo, no fue posible y no permitía seguir unos patrones que determinaran si era viable o no revisión. Por lo anterior, no se incluye este tipo de usuarios dentro de las bases de datos objetos de este estudio.

El método Adaboost es un método más estricto al momento de realizar el aprendizaje no supervisado y arroja usuarios puntuales y en menor cantidad para su revisión en terreno, permitiendo ahorrar en costos de revisión y garantizando una mayor efectividad; mientras que con Bagging se seleccionaba el 12.32 % de los usuarios, con Adaboost se concentraba en el 4.71% de los usuarios a revisar.

El método Adaboost presenta una mejora significativa en la efectividad respecto al Bagging ya que divide el universo de muestras con gran cantidad de clasificadores débiles y con características individuales que al unirse entregan un mejor clasificador que contiene las características de todos los clasificadores débiles, entregando un clasificador fuerte. El método Bagging utiliza muchos clasificadores individuales (Bootstrap) que al ser agregados entregan un clasificador promedio el cual suministra resultados aceptables, sin embargo, este método no tiene en cuenta las características individuales de cada uno de los clasificadores, siendo menos eficiente que Adaboost.

Las pruebas realizadas en terreno por parte de revisores de la EEP S.A. ESP fueron satisfactorias, ya que para la prueba del ciclo 10 de predios comerciales se obtuvo una efectividad del 23.08%, sobre el total de los predios intervenidos, con una proyección de

energía recuperada de 26.51 MWh y $13.13 millones de pesos. Para el ciclo 16 se obtuvo una efectividad del 33.33 %, con una proyección de energía recuperada de 5.7 MWh y $2.82 millones de pesos.

Este aplicativo permitirá al área de recuperación de energía de la EEP S.A. ESP realizar un análisis ciclo por ciclo, por tipo de servicio (comercial, industrial y residencial por estrato) y determinar campañas de revisión más pequeñas, aunque más efectivas, que permitirán disminuir costos en pago de revisiones, pero aumentar ingresos por Recuperación de Energía.

5.2. Trabajos a futuro

Para trabajos futuros, se sugiere implementar un tercer clasificador y crear un sistema de votación entre ellos (Baggin, Adaboost y el tercer método), con el fin de obtener una respuesta más específica. Se define además como irregular, cuando al menos dos de los tres clasificadores lo etiquetan como irregular.

Incluir las observaciones de lector en los métodos de entrenamiento, por ejemplo, en los casos que se tiene consumo 0, pero el predio se encuentra ocupado; esto con el fin de ampliar el universo para revisión.

Incluir en el aplicativo una variable que identifique en el período de evaluación y análisis de consumos, el mes exacto en el que se pueda evidenciar una diferencia significativa y enviar a revisión.

6. Bibliografía

[1]. Denice Jeanneth Romero López, Andrés Vargas Rojas. Modelo de Incentivos para la reducción de pérdidas de energía Eléctrica en Colombia. Revista Universidad Javeriana 13 Noviembre (2010).

[2]. Edison A.C. Aranha Neto, Jorge Coelho. Probabilistic methodology for Technical and Non-Technical Losses estimation in distribution system. Electric Power Systems Research 97 (2013) 93– 99.

[3]. Josif V. Spiric a, Miroslav B. Docic b, Slobodan S. Stankovic b. Fraud detection in registered electricity time series. Electrical Power and Energy Systems 71(2015) 42–5.

[4] Grupo de Reguladores de Energía y Gas de Europa. (2009). Treatment of losses by network operators (ERGEG Position Conclusions Paper). Bruselas: Autor.

[5]. Inigo Monedero A, Felix Biscarri a, Carlos Leon a, Juan I. Guerrero a, Jesus Biscarri b, Rocio Millan b. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. Electrical Power and Energy Systems 34 (2012) 90–98

[6]. X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, The Top 10 algorithms in data mining, V. K. Xindong Wu, Ed. Chapman & Hall/CRC, 2009.

[7]. J. R. Quinlan, C4.5, Programs for Machine Learning. Morgan Kaufmann Publishers Inc., 1993.

[8] J. Macqueen, \Some methods for classification and analysis of multivariate observations," in In 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[9] V. N. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., 1995.

[10] R. Agrawal and R. Srikant, \Fast algorithms for mining association rules in large databases," in Proceedings of the 20th International Conference on Very Large Data Bases, 1994.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, \Maximum likelihood from incomplete data via the em algorithm," in Journal of the Royal Statistical Society: Series B, 1977.

[12] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” Comput. Netw. ISDN Syst., vol. 30, pp. 107{117, 1998.

[13] D. Barber, Bayesian Reasoning and Machine Learning. Cambridge University Press, 2012.

[14] L. Fausett, Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Prentice-Hall, Inc., 1994.

[15] Bagging Predictors, 1996.

[16] Smith, T. (2004). “Electricity theft: a comparative analysis”. Energy Policy, vol. 32(18): 2067- 2076.

[17] Nagi, J.; Siah, K.; & Kiong, S. (2010). “Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines”. IEEE Transactions On Power Delivery, 25(2): 1162-1171.

[18] Davidson, I. (2002). “Evaluation and effective management of nontechnical losses in electrical power networks," Africa: IEEE AFRICON.

[19] Nagi, J. (2009). An intelligent system for detection of non-technical losses in tenaga national berhad (tnb) malaysia low voltage distribution network. Master's thesis, University Tenaga National, 2009.

[20] Suriyamongkol, D. (2002). Non-technical losses in electrical power systems. Estados Unidos: Ohio University.

[21] Berthold, M.; Borgelt, C.; Hppner, F.; & Klawonn, K. (2010). Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data. Springer Publishing Company, Incorporated, 1st edition,

[22] Han, J. & Kamber, Micheline (2005). Data Mining: Concepts and Techniques. 2a edición. Morgan Kaufmann Publishers: New York.

[23] Casa, N.; & Suncha, M. (2009). Control y reducción de pérdidas no técnicas de energía mediante el método balance de energía por transformador en 19 sectores de la provincia de Cotopaxi designados por ELEPCO S.A. Ecuador: LATACUNGA / UTC.

[24] Nizar, A.; Dong, Z.; & Wang, Y. (2008). Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems, 23(3), 946-955.

[25] Nizar, A.; Dong, Z.; Zhao, J.; & Zhang, P. (2007). A data mining based NTL analysis method. In Power Engineering Society General Meeting, 2007. IEEE (pp. 1-8). IEEE.

[26] Y. Freund and R. E. Schapire, \A decision-theoretic generalization of on-line learning and an application to boosting," in Proceedings of the Second European Conference on Computational Learning Theory-EuroCOLT '95, 1995.

[27] K. Sridharan and N. N. Schulz, \Outage management through amr systems using an intelligent data _lter," Power Delivery, IEEE Transactions on, vol. 16, pp. 669{675, 2001.

[28] E. Gontijo, A. Delaiba, E. Mazina, J. E. Cabral, J. O. P. Pinto et al., \Fraud identification in electricity company customers using decision tree," in Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004.

[29] A. H. S.V., Allera, “Load profiling for the energy trading and settlements in the uk electricity markets,” in DA/DSM Europe DistribuTECH Conference, 1996.

[30] L. I. Kuncheva and C. J. Whitaker, \Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Machine Learning, vol. 51, pp. 181-207, 2003.

[31] A. P. Birch and C. Ozveren, \An adaptive classi_cation for tari_ selection," in Metering Apparatus and Tari_s for Electricity Supply, 1992, Seventh International Conference on, 1992.

[32] S. Verdu, M. Garcia, F. Franco, N. Encinas, A. Marin, A. Molina, and E. Lazaro, “Characterization and identi_cation of electrical customers through the use of self-organizing maps and daily load parameters,” in Power Systems Conference and Exposition, 2004. IEEE PES, 2004.

[33] J. Jardini, C. M. V. Tahan, M. Gouvea, S. U. Ahn, and F. M. Figueiredo, “Daily load profiles for residential, commercial and industrial low voltage consumers,” Power Delivery, IEEE Transactions on, vol. 15, pp. 375-380, 2000.

[34] J.D. Gómez, “Aplicación de la metodología para evaluar el índice de potencialidad de infracción (ipi) en el mercado atenido por Dispac S.A E.S.P. en la ciudad de Quibdó-Chocó,” 2009, Universidad de los Andes-Bogotá, Colombia.

7. Anexos

Código:

En este proyecto se desarrolla una aplicación en el software Matlab, la cual contiene diferentes funciones para implementar las etapas de filtrado, asignación de etiquetas y clasificación, con el fin de cumplir con los objetivos propuestos en este documento.

Función principal:

A continuación, se presenta la función principal del proyecto. Esta función entrega una lista con las matrículas de los usuarios irregulares, a partir de una base de datos con el registro de los consumos, selección del método de clasificación, tipo de usuario y estrato. Esta función contiene diferentes sub-funciones que se encargan de realizar las etapas de filtrado, asignación de etiquetas y clasificación de los usuarios seleccionados, además entrega la lista de matrículas irregulares en un archivo de Excel.

```
function Irregularidades=Perdidas(baseDatos,metodo,uso,estra,file)
[Datos,datAti]=filtros(baseDatos,uso,estra);
if Datos==0
    Irregularidades=0;
    return
else
datos=Datos(:,2:13);
[Grupo_Sospechosos,Grupo_Normales,Datos_PRUEBA,placasPruebas]=AsignaEti
quetas(datos,Datos);
Datos=[placasPruebas,Datos_PRUEBA];
datos=Datos_PRUEBA;
Sospechosos=Grupo_Sospechosos;
Normales=Grupo_Normales;
red=Sospechosos;
blue=Normales;
```

```
datafeatures=[blue;red];
  figure(1),axis xy ;
  T=1:12;
  plot(T,mean(blue)','b',T,mean(red)','r');
  xlabel('Periodos [mes]');
  ylabel('Consumo [kW/h]');
  legend('Promedio Normales','Promedio Irregulares');
  title('Datos de Entrenamiento');
  figure(2),axis xy ;
  plot(T,datos,'k');
  title('Datos de Prueba');
  xlabel('Periodos [mes]');
  ylabel('Consumo [kW/h]');
if metodo==1 % AdaBoost
  dataclass(1:size(blue,1))=-1;
dataclass(size(blue,1)+1:size(blue,1)+size(red,1))=1;
  [classestimate,model]=adaboost('train',datafeatures,dataclass,50);
  blue=datafeatures(classestimate==-1,:);
red=datafeatures(classestimate==1,:);
  I=zeros(1e4,1e4);
  for i=1:length(model)
      if(model(i).dimension==1)
          if(model(i).direction==1), rec=[-80 -80 80+model(i).threshold
160];
          else rec=[model(i).threshold -80 80-model(i).threshold 160 ];
          end
      else
          if(model(i).direction==1), rec=[-80 -80 160
80+model(i).threshold];
          else rec=[-80 model(i).threshold 160 80-model(i).threshold];
          end
      end
      rec=round(rec);
      y=rec(1)+81:rec(1)+81+rec(3); x=rec(2)+81:rec(2)+81+rec(4);
      I=I-model(i).alpha; I(x,y)=I(x,y)+2*model(i).alpha;
  end
```

```
  testdata=datos;
  testclass=adaboost('apply',testdata,model);
  blue=testdata(testclass==-1,:);        red=testdata(testclass==1,:);
  Regular=Datos(testclass==-1,:); Irregular=Datos(testclass==1,:);
elseif metodo==2 % Bagging
  dataclass(1:size(blue,1))=0;
dataclass(size(blue,1)+1:size(blue,1)+size(red,1))=1;
  Datos_Train=datafeatures;
  Etiquetas=dataclass;
  Datos_Test=datos;
  [results,Resultado_BG] = Bagging( Datos_Train,Etiquetas,Datos_Test);
  blue=Datos_Test(Resultado_BG(:,3)==0,:);
red=Datos_Test(Resultado_BG(:,3)==1,:);
  Regular=Datos(Resultado_BG(:,3)==0,:);
Irregular=Datos(Resultado_BG(:,3)==1,:);
end

if isempty(blue)==1
    msgbox('Todos los clientes presentan posible irregularidad');
elseif isempty(red)==1
    msgbox('Nungún cliente presenta posible irregularidad');
else
  figure(3);
  plot(T,blue(:,:)','b',T,red(:,:)','r');
  xlabel('Periodos [mes]');
  ylabel('Consumo [kW/h]');
  legend('Normales');
if metodo==1
  title('Clasificación Datos de Prueba con el modelo AdaBoost');
elseif metodo==2
  title('Clasificación Datos de Prueba con el modelo Bagging');
end
end
if metodo==1
  Meto='AdaBoost';
elseif metodo==2
```

```
  Meto='Bagging';
end
% name=('MatriculasIrregulares',num2str(12),'.xlsx')
hora=fix(clock);
filename = ['Irregular_pruebaEEP_' Meto '_' date '_' num2str(hora(4))
'_' num2str(hora(5)) '_' num2str(hora(6)) ' ' file];

if isempty(red)==1
Irregularidades=0;
else
   if isempty(datAti)
   else
   datAti=datAti(:,1:13);
   Irregular=[datAti;Irregular];
   end
B(1:100,1)=' ';
xlswrite(filename,B);
xlswrite(filename,Irregular(:,1));
Irregularidades=Irregular;
figure(29)
bar([size(blue,1),size(red,1);0,0],'histc');axis([1,1.57,0,5+max(size(b
lue,1),size(red,1))]);grid on;
legend(['Normales:' num2str(size(blue,1))],['Irregulares:'
num2str(size(red,1))]);
end
end
```

Etapa de filtrado:

En la etapa de filtrado, se reemplazan los consumos faltantes por consumo igual a cero, luego se eliminan los usuarios con la mayoría de consumos iguales a cero; por último, se retiran los usuarios con consumos atípicos para ser incluidos en los usuarios irregulares estimados con el método de clasificación seleccionado. Esta función entrega los datos depurados y los usuarios con consumos atípicos.

```
function [datos,datAti]=filtros(baseDatos,uso,estra)
datAti=[];
[residencial,comercial,industrial,areas,oficial]=usuarios(baseDatos);%
Tipo de usuario
if uso==1
datosPre=residencial;
elseif uso==2
datosPre=comercial;
elseif uso==3
datosPre=Industrial;
end
[estrato1,estrato2,estrato3,estrato4,estrato5,estrato6]=estratos(datosP
re);% Estrato
if estra==1
datosPre=datosPre;
elseif estra==2
datosPre=estrato1;
elseif estra==3
datosPre=estrato2;
elseif estra==4
datosPre=estrato3;
elseif estra==5
datosPre=estrato4;
elseif estra==6
datosPre=estrato5;
```

```
elseif estra==7
datosPre=estrato6;
end
if size(datosPre,1)<10
respuesta=questdlg('Ha seleccionado un registro muy pequeño, pueden
producirse errores. ¿Desea continuar?','Registro muy
pequeño','Si','No','No');
else
respuesta=' ';
end
if strcmp(respuesta,'No')
   datos=0;
    return
else
CONSUMOS1=[datosPre(:,1) datosPre(:,5) datosPre(:,7) datosPre(:,9)
datosPre(:,11) datosPre(:,13) datosPre(:,15) datosPre(:,17)
datosPre(:,19) datosPre(:,21) datosPre(:,23) datosPre(:,25)
datosPre(:,27) datosPre(:,29) datosPre(:,31) datosPre(:,33)
datosPre(:,35)];
% Filtro1: Reemplaza o remueve datos faltantes: Este filtro se aplicó
en excel
CONSUMOS2=filtro2(CONSUMOS1);% Filtro2 Descarta consumos iguales a
cero, coincida con la observacion del lector....
[N_1,VM_1,DES_1,Vmax_1,Vmin_1]=Analisis(CONSUMOS2);
Muestras=1:(size(CONSUMOS2,2));
[datos,datAti]=filtro3(CONSUMOS2);% Filtro3 Elimina datos atípicos
% figure(20);
% plot(1:12,CONSUMOS1(:,2:13)')
% figure(21);
% plot(1:12,CONSUMOS2(:,2:13)')
% figure(22);
% plot(1:12,datos(:,2:13)')
end
```

Asignación de etiquetas:

En esta etapa se realiza el aprendizaje supervisado, el cual entrega una clasificación inicial de los usuarios de la base de datos depurada. Esta clasificación se divide en tres grupos de usuarios tales como; regulares, irregulares y datos de prueba. La metodología propuesta en este documento utiliza el criterio de la silueta para encontrar un agrupamiento adecuado de los usuarios de la base de datos depurada, de manera entregue tres grupos de usuarios a los métodos de aprendizaje no supervisado, para realizar las etapas de entrenamiento y clasificación final de los usuarios.

```
function
[Grupo_Sospechosos,Grupo_Normales,Datos_PRUEBA,placasPruebas]=AsignaEti
quetas(datos,Datos)
figure(10);
for k=2:1:10
IDX=kmeans(datos,k);
[S,H] = silhouette(datos,IDX);
silA(k)=mean(S);
end
[Val Pos]=max(silA);

[idx,ctrs]=kmeans(datos,Pos);
Datos_k=reshape(idx,length(datos(:,1)),1);
placas1=[];
placas2=[];
DemNV1=[];
DemNV2=[];
for i3=1:length(datos(:,1)) %size(datos(:,1),1)
Tmp1=length(find(Datos_k(i3,:)==1));
Tmp2=length(find(Datos_k(i3,:)==2));
May=[Tmp1 Tmp2];
Tm=find(May==max(May));
Usu(i3,1)=datos(i3,1);
Usu(i3,2)=Tm(end);
```

```
if Usu(i3,2)==1 %
DemNV1=[DemNV1;datos(i3,1:size(datos,2))];
placas1=[placas1;Datos(i3,1)];
elseif Usu(i3,2)==2
DemNV2=[DemNV2;datos(i3,1:size(datos,2))];
placas2=[placas2;Datos(i3,1)];
end
end
% Resultado analisis de agrupamiento
disp('Resultados de agrupamiento')
fprintf('Centro grupo 1: %3.1f\n',ctrs(1,:));
fprintf('Centro grupo 2: %3.1f\n',ctrs(2,:));
Aux1=DemNV1(:,2:end);
Aux2=DemNV2(:,2:end);
[Centro, Grupo]=min(ctrs);

placasS=[];
% Asignacion de registros de usuarios Normales y Sospechosos
Grupo_Normales=[];
Sospechosos=[];
if Grupo==1
Grupo_Normales=DemNV2;
Sospechosos=DemNV1;
placasS=placas1;
Etiqueta_N=ones(size(DemNV2,1),1);
else
Grupo_Normales=DemNV1;
Sospechosos=DemNV2;
placasS=placas2;
Etiqueta_N=ones(size(DemNV1,1),1);
end

Sospechosos_Vec=reshape(Sospechosos(:,1:size(Sospechosos,2)),numel(Sosp
echosos(:,1:size(Sospechosos,2))),1);
[idx_1,ctrs_1]=kmeans(Sospechosos_Vec,2);
Datos_k1=reshape(idx_1,length(Sospechosos(:,1)),12);
```

```
DemNV3=[];
DemNV4=[];
placas3=[];
placas4=[];
for i4=1:length(Sospechosos(:,1))
Tmp3=length(find(Datos_k1(i4,:)==1));
Tmp4=length(find(Datos_k1(i4,:)==2));
May=[Tmp3 Tmp4];
Tm1=find(May==max(May));
Usu1(i4,1)=Sospechosos(i4,1);
Usu1(i4,2)=Tm1(end);
if Usu1(i4,2)==1
DemNV3=[DemNV3;Sospechosos(i4,1:size(Sospechosos,2))];
placas3=[placas3;placasS(i4,1)];
elseif Usu(i4,2)==2
DemNV4=[DemNV4;Sospechosos(i4,1:size(Sospechosos,2))];
placas4=[placas4;placasS(i4,1)];
end
end
[Centro1, Grupo1]=min(ctrs_1);

placasPruebas=[];
Grupo_Sospechosos=[];
if Grupo1==1
Grupo_Sospechosos=DemNV3;
Etiqueta_S=zeros(size(DemNV3,1),1);
Datos_PRUEBA=DemNV4;
placasPruebas=placas4;
else
Grupo_Sospechosos=DemNV4;
Etiqueta_S=zeros(size(DemNV4,1),1);
Datos_PRUEBA=DemNV3;
placasPruebas=placas3;
end
```

Entrenamiento y clasificación:

A continuación, se presentan las dos funciones de los clasificadores AdaBosst y Bagging implementados en el desarrollo del proyecto.

Clasificador AdaBoost:

Este clasificador calcula una distribución inicial para los datos depurados, luego realiza un entrenamiento para varios clasificadores con distintos pesos que se calculan usando la medida del error. Finalmente se actualiza la distribución de cada clasificador y se obtienen las etiquetas de los usuarios de prueba. Este es un clasificador suave, que se vuelve robusto debido a que se ejecuta varias veces en la etapa de entrenamiento.

```
function
[estimateclasstotal,model]=adaboost(mode,datafeatures,dataclass_or_mode
l,itt)
switch(mode)
    case 'train'
        dataclass=dataclass_or_model(:);
        model=struct;
        D=ones(length(dataclass),1)/length(dataclass);
        estimateclasssum=zeros(size(dataclass));
        boundary=[min(datafeatures,[],1) max(datafeatures,[],1)];
        for t=1:itt
            [estimateclass,err,h] =
WeightedThresholdClassifier(datafeatures,dataclass,D);
            alpha=1/2 * log((1-err)/max(err,eps));
            model(t).alpha = alpha;
            model(t).dimension=h.dimension;
            model(t).threshold=h.threshold;
            model(t).direction=h.direction;
            model(t).boundary = boundary;
            D = D.* exp(-model(t).alpha.*dataclass.*estimateclass);
            D = D./sum(D);
```

```
            estimateclasssum=estimateclasssum
+estimateclass*model(t).alpha;
            estimateclasstotal=sign(estimateclasssum);

model(t).error=sum(estimateclasstotal~=dataclass)/length(dataclass);
            if(model(t).error==0), break; end
        end
    case 'apply'
        model=dataclass_or_model;
        if(length(model)>1);
            minb=model(1).boundary(1:end/2);
            maxb=model(1).boundary(end/2+1:end);
            datafeatures=bsxfun(@min,datafeatures,maxb);
            datafeatures=bsxfun(@max,datafeatures,minb);
        end
        estimateclasssum=zeros(size(datafeatures,1),1);
        for t=1:length(model);

estimateclasssum=estimateclasssum+model(t).alpha*ApplyClassTreshold(mod
el(t), datafeatures);
        end
        estimateclasstotal=sign(estimateclasssum);
    otherwise
        error('adaboost:inputs','unknown mode');
end

function [estimateclass,err,h] =
WeightedThresholdClassifier(datafeatures,dataclass,dataweight)
ntre=2e5;
r1=datafeatures(dataclass<0,:); w1=dataweight(dataclass<0);
r2=datafeatures(dataclass>0,:); w2=dataweight(dataclass>0);
minr=min(datafeatures,[],1)-1e-10; maxr=max(datafeatures,[],1)+1e-10;
p2c= ceil((bsxfun(@rdivide,bsxfun(@minus,r2,minr),(maxr-minr)))*(ntre-
1)+1+1e-9);   p2c(p2c>ntre)=ntre;
p1f=floor((bsxfun(@rdivide,bsxfun(@minus,r1,minr),(maxr-minr)))*(ntre-
1)+1-1e-9);  p1f(p1f<1)=1;
```

```
ndims=size(datafeatures,2);
i1=repmat(1:ndims,size(p1f,1),1);  i2=repmat(1:ndims,size(p2c,1),1);
h1f=accumarray([p1f(:) i1(:)],repmat(w1(:),ndims,1),[ntre ndims],[],0);
h2c=accumarray([p2c(:) i2(:)],repmat(w2(:),ndims,1),[ntre ndims],[],0);
h2ic=cumsum(h2c,1);
h1rf=cumsum(h1f(end:-1:1,:),1); h1rf=h1rf(end:-1:1,:);
e1a=h1rf+h2ic;
e2a=sum(dataweight)-e1a;
[err1a,ind1a]=min(e1a,[],1);  dim1a=(1:ndims); dir1a=ones(1,ndims);
[err2a,ind2a]=min(e2a,[],1);  dim2a=(1:ndims); dir2a=-ones(1,ndims);
A=[err1a(:),dim1a(:),dir1a(:),ind1a(:);err2a(:),dim2a(:),dir2a(:),ind2a
(:)];
[err,i]=min(A(:,1)); dim=A(i,2); dir=A(i,3); ind=A(i,4);
thresholds = linspace(minr(dim),maxr(dim),ntre);
thr=thresholds(ind);
h.dimension = dim;
h.threshold = thr;
h.direction = dir;
estimateclass=ApplyClassTreshold(h,datafeatures);

function y = ApplyClassTreshold(h, x)
if(h.direction == 1)
    y =  double(x(:,h.dimension) >= h.threshold);
else
    y =  double(x(:,h.dimension) < h.threshold);
end
y(y==0) = -1;
```

Clasificador Bagging:

Este método de clasificación crea una distribución de probabilidad empírica de los datos depurados, luego extrae una muestra aleatoria de los datos y calcula la distribución Bootstrap para la etapa de entrenamiento. Por último, se utilizan arboles de decisión para evaluar los datos de prueba con la distribución Bootstrap y entregar las etiquetas asignadas a los usuarios de prueba.

```
function [results,Resultado_BG] = Bagging(
Datos_Train,Etiquetas,Datos_Test)
feature_train_all =Datos_Train;
target_train_all=Etiquetas';
feature_test_all=Datos_Test;
target_test_all=[0];
num_samples_all =size(target_train_all,1);
num_bootstrap = 1;
num_samples_in_bootstrap =num_samples_all;
tree_bootstrap = cell(1,num_bootstrap);
for i_bootstrap = 1:num_bootstrap
id_random = randperm(num_samples_all);
id_select = id_random(1:num_samples_in_bootstrap);
feature_train_current = feature_train_all(id_select,:);
target_train_current = target_train_all(id_select,:);
tree_current =
classregtree(feature_train_current,target_train_current,'method','class
ification','categorical',1:12,'minleaf',1,'minparent',1);
tree_bootstrap{i_bootstrap} = tree_current;
results = cell(1,num_bootstrap);
for i_bootstrap = 1:num_bootstrap
result_current = eval(tree_current,feature_test_all);
results{i_bootstrap} = result_current;
end
C=results;
[BF,BC]=size(results);
for k=1:BC
```

```
A(:,k)=C{k};
end
for i=1:size(A,1)
for j=1:size(A,2)
AUX(i,j)=str2double(A{i,j});
end
end
DAN=transpose(AUX);
Final=zeros(3,size(DAN,2));
for l=1:size(DAN,2)
Final(1,l)=sum(DAN(:,l) == 1);
Final(2,l)=sum(DAN(:,l) == 0);
if Final(1,l)>Final(2,l)
Final(3,l)=1;
else
Final(3,l)=0;
end
end
Resultado_BG=transpose(Final);
end
```